kod szkolenia: DBX-DEA / ENG DL 3d

Databricks:Data Engineering Associate

This intensive three-day training guides participants through the full data engineer workflow on the Databricks platform. Combining content from Fundamentals, Lakehouse, and Data Transformation modules, participants will learn to import and explore data, build multi-stage pipelines in the Medallion architecture, and design advanced processing and automation using Delta Live Tables. The program is designed from the start to prepare for the Databricks Certified Data Engineer Associate exam, so it covers both basics of working with the Data Intelligence Platform and topics related to governance, data quality, and task orchestration.

Training recipients

– Data engineers preparing for the Data Engineer Associate exam.
– Data analysts and BI teams migrating processing to Databricks.
– Data scientists needing automated pipelines and clean, structured data.
– Professionals responsible for designing and standardizing data processing (lakehouse, medallion).

Benefits

– Full understanding of the Databricks Data Intelligence Platform: workspace, clusters, DBFS, Repos, and Workflows.
– Ability to load and clean data from multiple formats (CSV, JSON, Parquet) using SQL and PySpark.
– Understanding of the Lakehouse architecture and Medallion layers (Bronze, Silver, Gold), and pipeline design.
– Mastery of Delta Lake operations: ACID, time travel, merge, update, delete, vacuum, optimize, z-order.
– Building modular transformation pipelines using window functions, DML/DDL, and conditional logic.

– Automating processing with Delta Live Tables (DLT) and Databricks Workflows, including task parameterization.
– Introduction to Unity Catalog, Data Governance, and Delta Sharing to ensure security, access control, and data quality.

Training program

Day 1 – Fundamentals & Exploration
• Introduction to the platform: workspace, clusters, DBFS, Repos, Function registry (Repos)
• Data import from CSV, JSON, and Parquet: read.format, inferSchema, headers, delimiters; data types and casts
• Data exploration and visualization: display(), show(), describe(), summary(); generating statistics and plots in GUI
• Data cleaning: removing duplicates (dropDuplicates, distinct), handling nulls (dropna, fillna), casting and conditional transformations (when, otherwise)
• Basic transformations in SQL and PySpark: select, withColumn, filter, groupBy, orderBy, agg
• Creating temporary views and working with SQL in a notebook (createOrReplaceTempView)
• Introduction to Workflows: creating simple jobs, scheduling, and cyclic execution

Day 2 – Lakehouse & Delta Lake
• Lakehouse architecture and Medallion layers (Bronze → Silver → Gold); design strategies and best practices
• Delta Lake: ACID, schema enforcement and evolution, time travel, MERGE INTO, UPDATE, DELETE, CLONE
• Batch vs streaming load: COPY INTO, Auto Loader, structured streaming (readStream, writeStream), trigger times
• Building Bronze → Silver → Gold pipelines: raw data processing, cleaning, standardization, and aggregations
• DDL and DML operations on Delta tables; managed vs external tables
• Query optimization: partitioning, OPTIMIZE, ZORDER BY, VACUUM, execution plan analysis
• Introduction to Unity Catalog and basics of data management: catalogs, metastore, permissions, lineage

Day 3 – Data Transformation & Delta Live Tables
• Advanced transformations: creating binary, text, and numeric features; using conditional functions (case when, when/otherwise)
• Window functions: lag, lead, row_number, rank, rolling average; creating time-based and sequential features
• Modular pipelines in notebooks: code structure (load → transform → save), parameterization, and

running with dbutils.notebook.run()
• Silver → Gold processing: transformation, aggregation, and cleaning rules; reusability and separation of transformation logic
• Delta Live Tables (DLT): pipeline creation and declaration (CREATE LIVE TABLE), managing DAGs and schedules, parameterization, expectations as data quality tests
• Orchestration in Databricks Workflows: multi-task jobs, input parameters, alerts, and monitoring
• Data Governance & Quality: Unity Catalog (role-based access control, metastore), Delta Sharing, data lineage, cost considerations; audit log and quality test implementation
• Best practices: testing transformation code, notebook modularity, documentation, and repository versioning

Expected preparation of the participant

– Ability to work with data files (CSV/JSON) and basic knowledge of column/row/data type concepts.
– SQL and/or Python knowledge at a basic level.
– Experience with Databricks (at least 6 months) will make workshops easier but is not required.

Training Includes

- access to Altkom Akademia student

Training method:
– Lecture (30%) – introduction to platform, Lakehouse architecture, and data transformation concepts
– Workshops (30%) – pipeline design and implementation in notebooks and Workflows
– Exercises (40%) – hands-on tasks in SQL, PySpark, Delta Lake, Auto Loader, DLT

Main tools: Databricks Data Intelligence Platform, SQL, PySpark, Delta Lake, Auto Loader, Unity Catalog, Delta Live Tables, Databricks Workflows, notebooks
Note: During selected training sessions, the Academy provides Participants with access to paid tools and services necessary to complete the training program (in particular Azure, Fabric, and other cloud resources). The Participant is required to use the provided resources solely within the scope of the training agenda and in accordance with the trainer’s instructions. Use of these resources beyond the training scope—particularly for private projects, testing additional services, modifying the environment configuration, or undertaking activities that generate costs unrelated to the exercises—will result in the Participant being charged for costs exceeding the allocated limit, based on expenses incurred by the Academy.

Czas trwania

3 dni / 21 godzin

Language

Training: English

- Materials: English